Caring is Sharing – Exploiting the Value in Data for Health and Innovation
M. Hägglund et al. (Eds.)

doi:10.3233/SHTI230125

A Model for Multi-Institutional Clinical Data Repository

Karthik NATARAJAN [a,1], Chunhua WENG [a] and Soumitra SENGUPTA [a]
[a] *Columbia University, New York, NY, USA*
ORCiD ID: Karthik Natarajan https://orcid.org/0000-0002-9066-9431, Chunhua Weng https://orcid.org/0000-0002-9624-0214

Abstract. Creating a sustainable model for clinical data infrastructure requires the inclusion of key stakeholders, harmonization of their needs and constraints, integration with data governance considerations, conforming to FAIR principles while maintaining data safety and data quality, and maintaining financial health for contributing organizations and partners. This paper reflects on Columbia University's 30+ years of experiences in designing and developing clinical data infrastructure that synergizes both patient care and clinical research missions. We define the desiderata for a sustainable model and make recommendations of best practices to achieve a sustainable model.

Keywords. clinical data warehouse, data governance, data modeling

1. Introduction

Clinical data is essential to the learning health system. Since the 1980s, Columbia University's Department of Biomedical Informatics (DBMI) has been maintaining a clinical data repository [1], a homegrown web application connected to the repository for clinical care purposes, and a clinical data warehouse that has grown to include the longitudinal electronic health records (EHR) of >6 million patients to support institution wide clinical and translational research. Over the past 35+ years, this sustainable clinical data infrastructure has evolved and witnessed major changes in our organization: i.e., healthcare partners and facilities have grown through mergers and acquisitions, clinical systems have come and gone, and, more recently, our institution has converged on a single EHR system with our partner organizations. Currently, the care environment spans Columbia University Irving Medical Center (CUIMC), Weill-Cornell Medical Center, and NewYork-Presbyterian Hospital in New York City (and vicinity) in a tri-institutional Organized Health Care Arrangement with a single instance of EpicCare EHR for 11 hospitals, 2 faculty practice organizations, and a hospital medical group. In establishing a unified EHR, the leaders of the three institutions have set a clear direction for all, which is that the clinical data is to be equitably used for the success of care, research, finance, and education. This paper reflects on our key considerations and approaches to develop a sustainable data infrastructure and to achieve responsible and balanced tri-institutional management of clinical data for operations and research uses.

[1] Corresponding Author: kn2174@cumc.columbia.edu

2. Current Data Ecosystem

To overcome the challenge of data silos due to diverse data sources, nonintegrated data management systems with heterogeneous schemas, query languages, and APIs, a data lake approach was adopted as a viable solution for providing a schemaless repository for raw data with a common access interface. To date, our clinical data warehouse has been merged into an enterprise data lake consisting of both clinical and non-clinical data from the three institutions, with data feeds mostly in raw forms from EHRs, ancillary systems, and other transactional operational systems using a broad range of methods: HL7 messages, database replication, and batch extracts from source systems. Depending on the type of data, the lake stores data in form of SQL (for clinical data via Microsoft SQL server) and Non-SQL (for Telemetry and device data, via Hadoop from Cloudera) databases. Transformations and quality control checks are applied to the data lake to create curated databases, such as specific data marts (e.g., COVID-19 mart) or generic clinical data warehouses (with data standardized across EHRs using a local terminology system called Medical Entities Dictionary (MED) [2]), or operational datasets for dashboarding and reporting. In the past 5 years, Columbia has participated in or led national data networks such as OHDSI, PCORNet, National COVID Cohort Collaborative, and All of Us Research Program, requiring investment in maintaining heterogeneous common data models. With experience in harmonizing data from over 100 legacy systems, our enterprise data lake is compatible with these diverse data models. The terminology system and provenance mechanisms within the MED not only reconcile disparate data but also serves as a rich information source to understand the richness and limitations of the data. A key observation is that a data ecosystem is and will always be dynamic so that new sources or ETL (Extract, Transform, and Load) outputs should be incorporated constantly with system changes under a common governance.

3. Common Governance

Three institutions have collaboratively established two committees to govern data: the Committee creates policies such as a Data Sharing Agreement (currently 3rd generation in 10+ years) across all institutions, applies the policy towards the requirements for research, quality and operational requests, and sets rules about how access to data is provided and to whom while balancing the security requirements for data requests. For efficiency, routine operational (includes quality) requests are fast tracked, but research and cross-institutional requests are evaluated individually, examining IRB approvals, data use agreements or contracts related to external data sharing, and appropriate scoping. The committees are filled by multidisciplinary key stakeholders such as CIOs, CMIOs, research administration personnel, informatics leaders, leading clinician scientists, finance leaders, analytics leaders, researcher representatives, and clinical data engineers, from all three institutions. Incorporating researchers as stakeholders guarantees data access for responsible research personnel and creates educational opportunities for researchers to learn about the data and to generate more precise and efficient requests. Clinician scientists serve as effective thought leaders and influencers to emphasize to financial leadership that advanced research effectively contributes to exemplary clinical service.

4. Data Access: Cost and Expertise

The sustainability of a clinical data infrastructure and its services cannot be achieved without a fair cost model and a supportive mechanism for training and knowledge sharing. The cost for managing the enterprise data lake infrastructure is funded by tri-institutional operations groups, while data extraction is jointly funded by both operational analytics and research groups. We have previously reported the complexity of clinical data queries [3] and the iterative, human-centered nature of query clarification processes. We also found that use of self-service tool varies by experience and knowledge of users, which can potentially exacerbate the equity of data access. Therefore, we have explored data-driven methods for identifying common data elements needed by researchers [4], but mostly prioritized our effort towards manual service to aid researchers during data access. When a research request is approved, a set of data analysts, called Data Navigators (DN), extract data. DNs develop expertise about the data over time, and several technical approaches are used to disseminate such knowledge to support peer-based learning among DNs. The group of DNs conduct regular webinars on specific topics within the enterprise data lake. A Microsoft Teams group exist for DNs to communicate and exchange lessons learned on specific data. In addition, documentation exists in many different forms – wiki, data catalog, and repository of ETL code. DNs also help close the feedback loop by identifying data errors when they fulfill requests, assisting in making data more complete, and developing code and logic to query certain types of data that is shared across all DNs as well as operational analysts. Turnaround time is a key evaluation metric for DNs. The benefit of the DN model is that costs are shared across the spectrum of the institution, the department or the division, and the individual researcher based on what fits the need most. Each department can employ a data navigator to whom all data requests from that department are directed. Alternatively, a department can work with institutional IT to fund a DN partially and annually, or there can be a fee-for-service model for an individual researcher .

5. Desiderata for Sustainability

Our experience shows that sustaining the success of a clinical research infrastructure, specifically facilitating efficient access to clinical data for both operational and research uses, requires continual demonstration of the value of data and building trust in people and processes through transparency, fairness, partnership, and accountability, as further specified as the following desiderata:

5.1. Strong Partnership Among Stakeholders

All stakeholders must believe that disciplined and timely data availability is key for data driven insight for efficient health care operations, improvement of care quality, innovative clinical and informatics-based research, optimization of health finances, and overall success of the enterprise. All stakeholders must be committed to availability of data and responsible and secure use of data.

5.2. Extensible FAIR Data Infrastructure

The data infrastructure should follow the FAIR (Findable, Accessible, Interoperable, and Reusable) principle for data management. The data infrastructure has to accommodate

the establishment and cataloging of a myriad of data from different vendors and disparate databases with heterogeneous data models. Data ingestion should support diverse methods ranging from file transfer, remote queries, database replication, to incorporating HL7 (and now FHIR) based data feeds. Extensibility entails the ability to easily create different types of curated data marts for subsequent uses.

5.3. Comprehensive Data Governance

The platform and facilities to conduct both should be the same. The governance that establishes policy on how data is to be accessed and distributed must be a collaboration between operations and clinical administration to address the needs of both groups. The common governance is a trust-building activity, where operational, regulatory, and research interests are represented, and shared goals and results are emphasized.

5.4. A Cost-Sharing Model

Fair sharing of the costs of data consolidation, curation, extraction and analysis among the institution, the department, and the researcher is critical to sustain a collaborative research infrastructure. It is also an outcome of the trust models built under the common governance that includes appropriate representation of all stakeholders. Transparency in terms of use and discipline about how data requests are validated guide adherence to the institutional policies, benefiting all stakeholders.

5.5. Evidence of Value Add and Return on Investment

The request intake and fulfillment must be measured and reported to all stakeholders to monitor the efficiency of the system and individuals and identify areas for improvements as needed. Tracking research requests and connecting the requests to subsequent publications or grant awards is a concrete measure of return on investment. In addition, it is important to be able to measure success using survey techniques in collaboration with research administration. Transparency requires that each stakeholder is made aware of the metrics related to their investments.

5.6. Education and Technology Support

Comprehensive documentation and training are necessary in all aspects of the data infrastructure. A key requirement for success is development of personnel such as data engineers, analysts and scientists who understand the depth, nuances, and limitations of data. This is achieved by creating a combined educated workforce that collaborates and educates each other of new data resources.

5.7. Closed Feedback Loop and Support for Knowledge Sharing

It is desirable to close the feedback loop by engaging active contributions of different stakeholders, especially researchers and data engineers, who can report quality problems (or offer correction solutions) as they use the data, in collaboration with clinicians. A forum is needed to enable clinical data knowledge sharing among data users and other stakeholders.

6. Recommendations

In response to the aforementioned desiderata, we arrived at the following recommendations of best practices to improve the sustainability of clinical data infrastructure.

1. Inspire leadership to appreciate the full potential of clinical data for operations and research with latter informing how to improve content and processes.
2. Create an inclusive governance structure of all stakeholders that balances operational needs, security and privacy, and research needs without impeding progress.
3. Develop models of access but with appropriate controls for accountability and monitoring. Ensure that the controls are neither prohibitive nor lax and are set by a governance committee.
4. Identify and implement flexible cost sharing models that are reasonable based on the abilities of multi-level entities: e.g., institution, department, and researcher.
5. Track return of investment and value added and share this information with stakeholders.
6. Create solutions to educate the research community of the process, provide transparency of the process, and demonstrate accountability through metrics.
7. Develop talent with transferable knowledge to maintain continuity of services and provide tools for knowledge capture and exchange. Solicit active feedback from researchers and navigators, empowering them to play an active role in improving data quality.

7. Conclusion

Institutional leadership is critical for successful data infrastructure and effective analytics and research use of data. A collaborative, transparent model encourages proportional cost sharing and development of appropriate data expertise. Shared governance results in responsible sharing of data for secondary use, and in return, data quality is improved by continuous feedback from users. Since volume, variety, velocity, and veracity of health care data will only increase in the future, the infrastructure and use of data has to continually evolve. Strong governance, data engineering, and skilled data personnel are critical for continued success for future infrastructure developments, such as cloud computing, to support both research and operations.

References

[1] Johnson SB. Generic data modeling for clinical repositories [Journal Article]. J Am Med Inform Assoc. 1996;3(5):328-39. Available from: https://www.ncbi.nlm.nih.gov/pubmed/8880680.

[2] Cimino JJ, Johnson SB, Hripcsak G, Hill CL, Clayton PD. Managing vocabulary for a centralized clinical system [Journal Article]. Medinfo. 1995;8 Pt 1:117-20. Available from: https://www.ncbi.nlm.nih.gov/pubmed/8591133.

[3] Hruby GW, Rasmussen LV, Hanauer D, Patel VL, Cimino JJ, Weng C. A multi-site cognitive task analysis for biomedical query mediation [Journal Article]. Int J Med Inform. 2016;93:74-84. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27435950.

[4] Hruby GW, Hoxha J, Ravichandran PC, Mendonca EA, Hanauer DA, Weng C. A data-driven concept schema for defining clinical research data needs [Journal Article]. Int J Med Inform. 2016;91:1-9. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27185504.